DOCUMENT RESUME

ED 327 134 IR 014 722

AUTHOR Losee, Robert M., Jr.
TITLE Teaching Gateways and Bridges To Rank Broadcast Messages for Educational Networks.
PUB DATE 26 Aug 90
NOTE 17p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Coding; *Computer Networks; *Electronic Mail; Gateway Systems; *Information Management; *Information Networks; Man Machine Systems; *Mathematical Models; Relevance (Information Retrieval); Users (Information)
IDENTIFIERS *Ranking

ABSTRACT

Messages entering an educational information distribution network may be ranked for an ordered introduction into the network to maximize the timeliness of message arrivals over the set of users. Electronic mail, EDI documents, and broadcast news may be ranked by the users who choose to examine those messages of interest or benefit to themselves. Individuals choose to examine messages based on the message's features; a decision theoretic user interface may likewise examine features and determine which unexamined messages to present to the user. This interface may learn or estimate parameter values from user supplied relevance feedback. This knowledge may be sent to a gateway or bridge that similarly ranks messages based on their feature values and the knowledge the gateway has gained from the user supplied relevance feedback. Means to estimate parameters and the expected performance of the gateway ranking system can be computed algebraically. Future work on this approach to ranking of messages before they enter the network will involve further analyses of the efficacy of ranking. (17 references) (Author/DB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED327134

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Teaching Gateways and Bridges to Rank Broadcast Messages for Educational Networks

Robert M. Losee, Jr.
School of Information and Library Science
University of North Carolina
Chapel Hill, NC 27599-3360

losee@ils.unc.edu

August 26, 1990

Abstract

Messages entering an educational information distribution network may be ranked for an ordered introduction into the network to maximize the timeliness of message arrivals over the set of users. Electronic mail, EDI documents, and broadcast news may be ranked by the users who choose to examine those messages of interest or benefit to themselves. Individuals choose to examine messages based on the message's features; a decision theoretic user interface may likewise examine features and determine which unexamined messages to present to the user. This interface may learn or estimate parameter values from user supplied relevance feedback. This knowledge may be sent to a gateway or bridge that similarly ranks messages based on their feature values and the knowledge the gateway has gained from the user supplied relevance feedback. Means to estimate parameters at the gateway or

IR014722

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Robert M.Losee
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

the individual computer-human interface so as to support cooperative group efforts are examined. The expected performance of the gateway ranking system is given.

1 Introduction

Gateways and bridges provide interfaces between different networks in educational, as well as other environments. Those gateways and bridges having memory capacities large enough to store several messages may increase overall user satisfaction with the network if messages are distributed from the gateway in an order determined by the expected quality of the message. This ranking can result in the earlier arrival of more important messages than would occur if no ranking took place and the later arrival of the messages of lesser importance (Robertson, 1977).

All networks serve different users with different sets of needs. Users in elementary schools may be interested in material on bulletin boards at the same time that those in secondary schools are concerned about college admission requirements. Educational networks should reflect these localized interests and provide the information *likely* to be of greatest local interest first and those materials least likely to be of local interest last. In general, any network with store and forward capabilities at the entry point to the network should insert messages onto the network in the order of their expected importance to the users. This is not meant to encourage network censorship; although users in such a system indirectly choose which information is expedited and which is slowed down, all information is eventually broadcast, assuming the load presented to the network never exceeds its transmission capacity.

Gateways are normally thought of as connecting two different forms of networks, often using different communication protocols, while bridges connect networks using the same protocols. However, there is little functional difference between a bridge and a gateway that will have an impact on the

applicability of the following analysis. In addition, little or no filtering based on message content is performed by most existing networks or bridges, but we refer to these proposed modified gateways and bridges as *gateways* to sim;lify nomenclature.

A pair of ranking mechanisms is proposed here, both of which have the capability to learn characteristics of messages ranked high or low. One ranking of messages is provided by the user's computer immediately before the user examines messages. This ranking's function is to present the message most likely to be of interest to the user first, with messages then being presented in decreasing order of expected interest. When the user provides relevance feedback, an indication that a message has been or has not been found of interest to the user (relevant or nonrelevant), the ranking algorithm may "learn" the characteristics of those messages that the user finds relevant or nonrelevant.

A network gateway may similarly learn which messages are of greatest interest to those users on it's network, or other networks downstream of the network of concern, where *downstream* sites are those that receive broadcasts from the gateway. The feedback provided by an individual user to his or her computer may be transmitted, in summary form, to the network gateway. The gateway may then learn the characteristics of messages of interest to its' users, so that it may provide more efficient and timely distribution of messages.

Broadcasting one's feedback effectively discloses the characteristics of messages that one finds relevant, providing little privacy if others can monitor feedback messages transmitted toward the gateway or bridge. Several solutions provide a level of user privacy. One may encrypt the feedback sent to the gateway using a public-key private-key cryptographic system. The user encodes the feedback by using an algorithm and the gateway's public key. The gateway may then decode the message using its' private-key. A second, weaker, solution is to transmit feedback anonymously, using unsigned

or false "from" addresses on messages. The weakness of this methods is that the physical origin of a signal can be located by the determined snoop. A third means of ensuring privacy is to have the message filter operate on only the user's machine (Gifford, 1990, Losee, 1989, Malone et al., 1987) and not broadcast any information concerning the interests of the users.

Both the ranking at the user's computer and the ranking by the gateway assume the availability of all message features for examination by the network gateway or the user's computer. Although the models developed here are inherently media independent, of greatest concern at present is the economical and timely transmission of textual material, written in a natural language and capable of being stored in digital form. Message features might include such attributes as the name of the sender, the site of the sender, the date the message was sent, and whether the message contains the word "network". By learning the features in those messages found to be of interest to the users, this system may rank those messages containing features often found in documents of interest ahead of documents containing no features held in common with the set of documents of interest. Other features may be derived from a process such as factor analysis, which can isolate a set of statistically independent factors underlying a set of dependent features.

A *virtual feature* provides state dependent information about a message and the reader (Losee, 1989). Such a feature might be "shares feature X with the message currently being examined." Such a virtual feature is dependent upon the state of the system at the time that the message is being considered for examination. If the user is reading a message on TCP/IP, the user may find another message on the same topic to be of greatest interest him or her at that point in time. This emulates the ability of some news reading programs to allow the user to read through all messages dealing with a particular subject, before going on to those dealing with other topics. The ranking procedure proposed here can learn of the user's preference for this form of ranking and integrate it into other preferences expressed by the user through

the relevance feedback provided.

Most other message filtering systems have required that the characteristics of desirable messages, or similarly, characteristics of those messages to be excluded from consideration, be *programmed* into the filter, either as part of an expert system rule (Malone et al., 1987) or as part of an explicit filter (Gifford, 1990). While these filters have achieved some success, it is likely that an adaptive filter such as that described here provides superior performance because it (1) requires no prior "programming knowledge" or computer literacy to use, (2) can precisely learn the user's interests, and (3) can change over time as the user's interests vary.

2 Deciding to Examine Messages

A user should choose to examine an incoming message if the cost of examining the message is less than the cost of not examining the message. When the user does not know before hand the cost of examining or not examining the message, the user should base the decision on the expected cost of examining the message; that is, a message should be examined if the expected cost of examining the message is less than the expected cost of not examining the message. The expected cost of examining the message may be estimated as $\Pr(g|m)C_{e,g} + \Pr(\bar{g}|m)C_{e,\bar{g}}$, where g and $\bar{g}$ represent good and not good messages, sometimes referred to as relevant and nonrelevant messages, e and $\bar{e}$ represent the decision to examine and not examine a message, m represents the message with all it's features, and $C_{e,g}$ represents the conceptual cost of performing action e (examine) when the message is of class g (good). The probability that a message is good or not good, given the message characteristics, serves as a weighting factor for the two cost constants.

Through algebraic manipulation, the following decision rule may be de-

rived: choose to examine a message if

$$\frac{\Pr(g|m)}{\Pr(\bar{g}|m)} > \frac{C_{e,\bar{g}} - C_{\bar{e},\bar{g}}}{C_{\bar{e},g} - C_{e,g}} = Constant,$$

where *Constant* denotes a cost ratio combining the earlier cost figures.

The cost constant may be difficult to estimate for a given message, making the use of the above expression difficult. Users may avoid supplying this cost constant by ranking messages by the ratio of the probabilities on the left. This ranking will place those messages with the highest ratio to be retrieved first, while those with the lowest ratio will be examined last.

By using Baye's Rule, and assuming the statistical independence of features, the decision rule may be reexpressed as

$$\left(\prod_{i=1}^{j} \frac{\Pr(m_i|g)}{\Pr(m_i|\bar{g})} \right) \frac{\Pr(g)}{Pr(\bar{g})},$$

assuming the availability of j features concerning the message. The general performance of a decision procedure based on such a ranking procedure is discussed in (Losee, 1987, Raghavan et al., 1989).

If the expression, $\Pr(g)/Pr(\bar{g})$ is dropped from this expression because it is constant for all messages, the decision rule may still be used to rank messages. By other manipulations described by (Losee, 1988), the expression may be modified to

$$\prod_{i=1}^{n} \frac{\Pr(m_i|g)/Pr(0_i|g)}{\Pr(m_i|\bar{g})/Pr(0_i|\bar{g})},$$

if this value is only to be used for the ranking of messages. The incorporation of the probability that a feature has the value 0 performs a normalizing function, of particular value when nonbinary feature distributions are used. The binary distribution is used in this paper, but we keep the 0 feature value normalization to allow for future generalization.

3 Binary Independent Features

The application of this user-based message ranking procedure requires that the probabilities of independent message features be estimated. One relatively simple model that has proved effective in information retrieval (Losee et al., 1986) is the assumption of binary independent features. Such features have the value either 0 or 1, indicating the presence or absence of the feature in a particular message. A message might, for example, contain the word *gateway* (1) or not contain the word *gateway* (0). Other distributions may be used to describe features, but this nonbinary information may be expressed in binary form with some loss of information. For example, English language words appear to be roughly Poisson distributed; these terms may be described by a binary distribution which assigns the value 1 if the term is present and 0 if the term is absent. A binary feature's distribution in the class of relevant messages may be described as $\Pr(m_i|p_i) = p_i^{m_i}(1-p_i)^{1-m_i}$, with a similar expression for nonrelevant messages using the parameter q_i in place of the parameter p_i.

The parameter p_i denotes here the probability that a feature will have a value of 1 given that message i is relevant, while the parameter q_i represents the probability that a feature will have a value of 1 given that message i is nonrelevant.

Applying the binary independence feature assumption to our earlier decision rule suggests that documents now be ranked by

$$\prod_{i=1}^{n} \left(\frac{p_i/(1-p_i)}{q_i/(1-q_i)} \right)^{m_i}.$$

To simplify the understanding of this expression, it may be logarithmically transformed to

$$MSV = \sum_{i=1}^{n} m_i \log \frac{p_i/(1-p_i)}{q_i/(1-q_i)}.$$

A positive Message Status Value (MSV) indicates the ability of the features of the message to discriminate well between relevant and nonrelevant messages. Those messages with an MSV near 0 have features equally likely to occur in relevant and nonrelevant messages. One can generalize from this expression to say

$$MSV = \sum_i m_i \Delta_i,$$

where the Δ_i component measures the degree to which feature i distinguishes between messages that should be examined because they're relevant and messages that should not be examined because they are presumably nonrelevant. In the binary independence model that we have proposed

$$\Delta_i = \log \frac{p_i/(1-p_i)}{q_i/(1-q_i)},$$

represents the discrimination value of a binary independently distributed feature.

4 Relevance Feedback

For this message ranking system to improve performance for the individual as time progresses, it must incorporate the knowledge gained from relevance feedback. The system may then revise its estimates of parameter values, e.g., p_i and q_i, based on Bayesian principles incorporating existing knowledge with newly received knowledge. If parameter p_i of the binary distribution is assumed to be distributed as described by the beta distribution, p_i may be estimated as $(B_i+b_i)/(C_i+c_i)$, where B_i and C_i are arbitrarily chosen values such that $B_i < C_i$, b_i denotes the number of relevant examined messages containing the ith feature and c_i denotes the number of relevant messages examined. These lower case values represent the relevance feedback. B_i and C_i represent knowledge available prior to examining any relevant messages

and b_i and c_i represent posterior or learned knowledge obtained from user supplied relevance feedback judgements. A similar set of parameters exist for estimating q_i.

If no documents have been retrieved yet, parameters b_i and c_i are both 0. If parameters B_i and C_i are arbitrarily assigned the values 1 and 2 respectively, the initial estimate of p_i is 1/2. If a message was examined and it was found relevant and it contained feature i, b_i would become 1 and c_i would become 1. The estimate of p_i then increases to $(1+1)/(2+1) = 2/3$. By continuing to supply relevance feedback, the estimates of p_i and q_i are refined. The initial estimate of p_i, (B_i/C_i), can be overridden by a large amount of relevance feedback. However, the initial estimate of p_i is of importance because of the need to initially retrieve some relevant messages and to exclude some nonrelevant messages in order to obtain some properly classified messages for the relevance feedback.

An initial estimate of p_i in which the user has a certain degree of confidence may be represented by setting B_i to the estimate multiplied by the number of messages that would have to be examined to yield this degree of confidence and setting C_i to the number of messages that would have to be examined to yield the same degree of confidence. If parameter p_i were to be initially estimated as 1/2, for example, and this is to be held with the confidence that would be obtained from examining 10 documents, one would set B_i to $1/2 \times 10 = 5$ while C_i is set to 10.

5 Gateway Operation and Adaptation

Gateways and bridges function as interfaces between different networks. When information travels from one network to another network, the gateway may store the information as it leaves one network, then forward it onto the second network at a later time. Before the information is retransmitted from the first network to the second network, it may be manipulated so that the

information is transmitted in a ranked order, placing the message most likely to be found interesting first, that message next most likely to be found interesting second, and so forth. This ranking is similar to the ranking performed for individual users by individual computers as discussed above.

If the gateway is to provide a ranking service, it needs to be able to estimate parameter values. These parameters may be estimated at the gateway in a manner similar to that used at individual user sites. We denote all parameters of the gateway by using the prime ($'$) symbol.

As with the single user model of the decision process, a gateway may decide to forward messages iff

$$\frac{\Pr(g|m)'}{\Pr(\bar{g}|m)'} > \frac{C'_{e,\bar{g}} - C'_{\bar{e},\bar{g}}}{C'_{\bar{e},g} - C'_{e,g}} = Constant',$$

where *Constant'* denotes a cost ratio combining the primed cost figures. The probability distributions $\Pr(g|m)'$ and $\Pr(\bar{g}|m)'$ can be interpreted here as the joint distributions of the probabilities of each user on the "downstream" side of the gateway. All users have their own set of discrimination values representing their relative interests in individual characteristics occurring in relevant messages. These are combined by the gateway, yielding a discrimination value for the network.

Messages may be ranked by their network message status value,

$$MSV' = \sum_{i=1}^{n} m_i \log \frac{p'_i/(1-p'_i)}{q'_i/(1-q'_i)}.$$

Parameters p'_i and q'_i may be estimated as with the user's ranking system, that is, based on a set of initial parameters B'_i and C'_i as well as from user supplied feedback in the form of b'_i and c'_i values. Here b'_i represents the total number of relevant messages examined by users that had the feature and c_i represents the number of relevant messages examined by users. As before B'_i and C'_i represent initial knowledge and may be estimated using a simple

technique such as equating $B_i'/C_i' = 1/2$. The numbers are obtained from individual sites, which report on the number of relevant and nonrelevant messages read and the number considered relevant or nonrelevant with a particular feature.

The USENET currently compiles statistics representing user interest and readership in news groups. This is a form of relevance feedback. These statistics provide a limited amount of information about user's interests. If an examination of a message in a news group may be interpreted as indicating that the message was found relevant then the USENET's statistics can be interpreted as providing relevance feedback about a sing! feature, the name of the news group. Expanding the feedback to incorporate other features should increase system performance by transmitting messages of greater importance before those of lesser importance. As the volume of broadcast news traffic and the number of users increases, the ability to prioritize messages grows in importance.

Systems supporting cooperative work efforts may provide feedback to the gateway in a slightly different form. All personnel on the group project wishing to receive group related materials may choose a "leader" who supplies feedback to the gateway that is meant to represent the entire group's feedback. Such feedback includes a multiplier n, indicating that this feedback represents the interests of n people on the project. This group feedback can reduce network traffic by cutting back on the number of users transmitting feedback from workstations to the gateway.

The present form of broadcasting found on most networks is one implicitly consistent with all users representing a single large cooperative group with no special interests. If feedback were to be provided by a single individual or program, e.g. an expert system filter such as that used in Malone's *Information Lens*, the gateway would function as though there were a single, cooperative group providing some feedback. If the gateway functioned as suggested in this paper, with feedback coming from each user, the gateway

optimizes broadcast performance for the individual user.

6 Performance

The performance of a gateway ranking system may be computed as follows. The number of users downstream of the gateway is denoted as u and the number of ranked messages as N, while m_r represents the message of rank r, numbered from 1 to N, and where 1 is the first transmitted and N the last transmitted. The variable t is the time it takes to transmit an average message on the network.

The performance of a ranking gateway can be understood as the decrease in time it takes for relevant messages to enter the network and the increase in time it takes for nonrelevant messages to enter the network. The decrease in time for relevant documents may be computed as

$$T_{Rel} = tu \sum_{r=1}^{N} \left[\Pr(Rel|m_r) \left(\frac{N+1}{2} - r \right) \right]$$

and the decrease for nonrelevant messages as

$$T_{Nonrel} = tu \sum_{r=1}^{N} \left[(1 - \Pr(Rel|m_r)) \left(\frac{N+1}{2} - r \right) \right].$$

A traditional store and forward gateway that doesn't rank messages decreases message transmission time by $T_{Rel,Random}$ and increases it by $T_{Nonrel,Random}$.

The performance of a system that ranks messages may be similarly determined as the decrease in transmission time of relevant messages, $T_{Rel,Ranked}$ minus the increase in transmission speed of nonrelevant messages, $T_{Nonrel,Ranked}$.

As an example, assume that $N = 5$ with two messages with $\Pr(Rel) = .8$, with the remaining 3 messages having $\Pr(Rel) = .3$. In "random order," they are arranged so that the first and last are $\Pr(Rel) = .8$, while when "ranked,"

the first two messages are $\Pr(Rel) = .8$. Assume that $u = 1000$ and that t equals 1 millisecond.

$$\begin{aligned} T_{Rel,Random} &= 0.001 \times 1000[.8(3-1) + .3(3-2) + .3(3-3) \\ &\quad +.3(3-4) + .8(3-5)] \\ &= 0. \end{aligned}$$

$T_{Nonrel,Random}$ similarly equals 0. A random ranking neither increases or decreases the time it takes to transmit messages through the gateway.

The ranking of messages does result in an expected decrease in transmission time for relevant messages,

$$\begin{aligned} T_{Rel,Random} &= 0.001 \times 1000[.8(3-1) + .8(3-2) + .3(3-3) \\ &\quad +.3(3-4) + .3(3-5)] \\ &= 1.5 \; sec. \end{aligned}$$

and an expected increase in time for nonrelevant messages,

$$\begin{aligned} T_{Rel,Random} &= 0.001 \times 1000[.2(3-1) + .2(3-2) + .7(3-3) \\ &\quad +.7(3-4) + .7(3-5)] \\ &= -1.5 \; sec. \end{aligned}$$

The marginal increase in performance obtained by using a ranking gateway can be computed as the difference of the increased transmission timeliness with and without ranking, or

$$(T_{Rel,Ranked} - T_{Nonrel,Ranked}) - (T_{Rel,Random} - T_{Nonrel,Random}).$$

The overall increase using the example above becomes $(1.5 - (-1.5)) - (0 - 0) = 3$ seconds.

7 Summary

The presentation of messages in a ranked order can increase user satisfaction with the network. Ranking can be provided by either the computer that the user is sitting at or a network gateway. The proposed model allows for the incorporation of relevance feedback into the user interface, allowing the interface to learn the characteristics of messages that the user finds of interest. This further allows the system to adapt to the user.

Gateways may use this knowledge provided by individual computer sites in learning the characteristics of the messages that the users on the network find of greatest interest. If the gateway ranks messages based upon a decision theoretic principle, it can incorporate feedback in a way consistent with formal models. This feedback is already provided by users of individual news reading programs that are consistent with decision theoretic principles.

Future work on this approach to ranking of messages before they enter the network will involve further analyses of the efficacy of ranking . A second area of concern is whether this model may be used simultaneously with different classes of messages, such as real time, electronic mail, or messages of relatively low importance.

References

Bookstein, A. (1983a). Information retrieval: A sequential learning process. *Journal of the American Society for Information Science*, 34(4):331–342.

Bookstein, A. (1983b). Outline of a general probabilistic retrieval model. *Journal of Documentation*, 39:63–72.

Chang, S. K. and Leung, L. (1987). A knowledge based message management system. *ACM Transactions on Office Information Systems*, 5(3):213–236.

Croft, W. B. and Harper, D. (1979). Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35(4):285–295.

Denning, P. (1982). Electronic junk. *Communications of the ACM*, 23:163–165.

Gifford, D. K. (1990). Polychannel systems for mass digital communication. *Communications of the ACM*, 33(2):141–151.

Hiltz, S. R. and Turoff, M. (1985). Structuring computer-mediated communications systems to avoid information overload. *Communications of the ACM*, 28(7):680–689.

Kerr, E. B. and Hiltz, S. R. (1982). *Computer Mediated Communication Systems*. Academic Press, New York.

Losee, Jr., R. M. (1987). Predicting document retrieval system performance using an expected precision measure. *Information Processing and Management*, 23(6):529–537.

Losee, Jr., R. M. (1988). Parameter estimation for probabilistic document retrieval models. *Journal of the American Society for Information Science*, 39(1):8–16.

Losee, Jr., R. M. (1989). Minimizing information overload: The ranking of electronic messages. *Journal of Information Science*, 15(3):179–189.

Losee, Jr., R. M., Bookstein, A., and Yu, C. T. (1986). Probabilistic models for document retrieval: A comparison of performance on experimental and synthetic databases. In *ACM Conference on Research and Development in Information Retrieval*, pages 258–264.

Malone, T. W., Grant, K. R., Turbak, F. A., Brobst, S. A., and Cohen, M. D. (1987). Intelligent information sharing systems. *Communications of the ACM*, 30:390–402.

Raghavan, V. V., Bollmann, P., and Jung, G. S. (1989). Retrieval system evaluation using recall and precision: Problems and answers. In *ACM Annual Conference on Research and Development in Information Retrieval*, pages 59–68, New York. ACM Press.

Robertson, S. E. (1977). The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304.

Robertson, S. E., Van Rijsbergen, C. J., and Porter, M. (1981). Probabilistic models of indexing and searching. In Oddy, R., Robertson, S. E., van Rijsbergen, C. J., and Williams, P. W., editors, *Information Retrieval Research*, pages 35–56, London. Butterworths.

Smith, H. T., editor (1984). *Computer Based Message Systems*. North-Holland, Amsterdam.